Análisis topológico de señales ECG

Sergio Alejandro Esparza González - A01625430 - IDM,
Instituto Tecnológico y de Estudios Superiores de Monterrey,
1 de mayo de 2025

Resumen

Este trabajo presenta un enfoque combinado de análisis topológico de datos (TDA) y aprendizaje automático para la clasificación binaria de latidos cardíacos en ECG. Primero, cada latido se incrusta en un espacio de fases mediante encaje sliding-window, determinando el retardo óptimo con AMI y la dimensión de incrustación con SVD. Después, se extraen sus invariantes topológicos a través de homología persistente resumida en paisajes de persistencia (Persistent Landscapes). Luego, un clasificador de bosque aleatorio (Random Forest) entrenado sobre estos vectores topológicos consigue un accuracy de 0.88 y un AUC de 0.94, demostrando alta sensibilidad para detectar arritmias (recall 0.88, precision 0.94), y preservando un bajo nivel de falsos positivos. Finalmente, se compara con un modelo idéntico sobre datos crudos y se observa que, aunque el enfoque TDA no supera por completo al baseline, aporta una representación más interpretable y robusta de la morfología del latido, con potencial para mejorar sistemas de apoyo al diagnóstico clínico.

I. Introducción

Las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad a nivel mundial, superando incluso a ciertos tipos de cáncer y enfermedades respiratorias. La detección temprana y el diagnóstico previo de estas afecciones son un desafío, pues se involucran diversos factores causales, de riesgo y signos de las enfermedades. En este contexto, el electrocardiograma (ECG) es una herramienta de diagnóstico fundamental, pues registra la actividad eléctrica del corazón, permitiendo identificar arritmias, isquemia, miocardia y otras patologías cardíacas. No obstante, la interpretación manual de ECG es un proceso especializado y laborioso; por ello, en años recientes ha cobrado gran importancia el análisis automático de ECG mediante algoritmos de inteligencia artificial y aprendizaje automático, con el fin de asistir en el diagnóstico de forma rápida y consciente (Hernández-Lemus et al., 2024).

Un ECG es un registro temporal de la actividad eléctrica del corazón, reflejando el ciclo de despolarización y repolarización de las aurículas y ventrículos cardíacos. Cada latido normal produce impulsos eléctricos en un patrón de ondas característico, asociado a eventos específicos del ciclo cardíaco (Cascino & Shea, 2023). El ECG captura estos impulsos para analizar el ritmo y la función del corazón. Estos componentes se registran mediante electrodos en la superficie del cuerpo, aprovechando que los tejidos corporales conducen la corriente eléctrica generada por el corazón. Un ECG estándar usa múltiples derivaciones (configuraciones de electrodos) para obtener perspectivas distintas de la actividad cardíaca.

El PTB Diagnostic ECG Database (PTB-DB) es un conjunto de datos (dataset) público, ampliamente usado para la investigación en diagnóstico cardíaco automatizado (PhysioNet, 2004). Este fue recopilado por la institución Physikalisch-Technische Bundesanstalt (PTB) de Alemania junto con clínicas cardiológicas, y luego puesto a disposición en PhysioNet, y en este caso se utilizó la versión disponible en Kaggle (Fazeli, 2018).

I-A. Preguntas de investigación

A partir de la problemática descrita, se plantean las siguientes preguntas de investigación:

1. ¿Existen diferencias topológicas claras entre latidos normales y latidos anormales en señales ECG?
2. ¿Los vectores de características basados en TDA, por ejemplo, los Persistence Landscapes permiten clasificar latidos normales vs. anormales con una alta precisión?
3. ¿El uso de algoritmos de Geometría y Topología brinda una ventaja significativa frente a los enfoques convencionales de aprendizaje automático en la clasificación de latidos?

I-B. Literatura relevante

En los últimos años, los enfoques basados en redes neuronales profundas, particularmente las CNN 1-D y los modelos LSTM, han alcanzado desempeños comparables a cardiólogos humanos en la detección de arritmias complejas (Hannun et al., 2019). Estas redes aprenden representaciones directamente de la forma de onda, omitiendo la extracción manual de características, tal como lo hacen los métodos más clásicos, como SVM o árboles de decisión sobre características diseñadas manualmente (intervalos temporales, amplitudes, coeficientes de ondaleta). Estos métodos han logrado resultados aceptables (de Chazal et al., 2004), pero normalmente quedan por debajo de las redes profundas en grandes datasets.

Al mismo tiempo, ha emergido un paradigma alternativo: la geometría y la topología aplicadas a la ciencia de datos. El Análisis Topológico de Datos (TDA) proporciona herramientas para extraer invariantes de la ”forma” en espacios de alta dimensión. En señales temporales, el TDA captura propiedades geométricas del latido más allá de las características convencionales unidimensionales. Una técnica clave es el sliding-window embedding, que reconstruye el espacio de fases de la señal. Sobre la nube resultante se computa la homología persistente, identificando conectividad y bucles a múltiples escalas (Perea & Harer, 2015).

El resultado de la homología persistente se puede resumir diagramas de persistencia, que pueden vectorizarse mediante Persistence Landscapes, lo que posibilita alimentar a distintos clasificadores. Estudios preliminares sugieren que las características basadas en TDA pueden capturar diferencias sutiles en la forma del ECG, que pudieran ser indicativas de patología. Por ejemplo, se ha aplicado TDA para distinguir los patrones en la variabilidad cardíaca para la detección de arritmias mediante combinaciones de diagramas de persistencia con redes neuronales modulares, logrando buenos resultados (Chung et al., 2021), y especialmente el uso de los paisajes de persistencia o Persistence Landscapes se ve prometedor para combinarse con algoritmos de aprendizaje automático.

En conjunto, la literatura sugiere que las características topológicas capturan diferencias morfológicas sutiles entre latidos normales y anormales, y que pueden igualar o superar a los enfoques profundos cuando los datos son limitados o cuando se busca interpretabilidad adicional.

II. Objetivo e hipótesis

II-A. Objetivo

Desarrollar y evaluar un sistema de clasificación binaria de latidos normales y anormales, aprovechando los descriptores topológicos derivados de la homología persistente, tal como los Persistence Landscapes, obtenidos a partir de la incrustación de ventana deslizante de la señal ECG, y comparar su desempeño de manera cuantitativa con enfoques de aprendizaje profundo y con clasificadores clásicos que utilizan características manuales, como los bosques aleatorios (Random Forests).

II-B. Hipótesis

H1. *Diferencia topológica*. Los paisajes de persistencia difieren de forma significativa entre latidos normales y anormales, revelando la presencia de patrones geométricos característicos de cada clase.

H2. *Capacidad predictiva del TDA*. Un clasificador Random Forest entrenado exclusivamente con los vectores de paisajes alcanza un rendimiento elevado (precisión global y AUROC $\geq 0{,}90$) al clasificar los latidos en el dataset.

H3. *Ventaja comparativa*. El modelo basado en descriptores topológicos supera en al menos 10 % de AUROC a un modelo que no utiliza TDA, mostrando que la información geométrica aporta valor añadido a la tarea de clasificación.

III. Metodología

III-A. Datos y preprocesamiento

Se emplea el conjunto de datos dado por PTB Diagnostic ECG Database obtenido en Kaggle. Aquí, los datos se dividían en dos archivos de tipo CSV: 'ptbdb_normal.csv' y 'ptbdb_abnormal.csv', donde se tienen registros de latidos de pacientes sanos y con anomalías cardíacas respectivamente. Entre los dos datasets se tienen 14552 muestras de latidos distintas (4046 y 10506 respectivamente), tomadas a una frecuencia de 125 Hz, dado que un latido dura aproximadamente 1.5 segundos, se tomaron así 187 mediciones de voltaje de ECG, pero si un latido termina en menos de las 187 mediciones, se añaden ceros al final para forzar una longitud común. Cabe destacar que las mediciones están escaladas con min-max en el intervalo de 0 a 1.

Entonces, se tienen 187 variables correspondientes a las mediciones de ECG, y una al final que corresponde a una de las dos clases: pacientes sanos (0) o pacientes con alguna anomalía (1), según sea el dataset.

Para preprocesar los datos, primero se comprobó que no hubiera datos nulos, y afortunadamente no hubo en ningún dataset. Después se eliminaron los registros duplicados, en donde se eliminó uno del dataset de latidos normales, y 6 del dataset de latidos anormales, quedando 4045 y 10500 registros de manera respectiva. Finalmente se unió el dataset a uno solo de 14545 registros.

III-B. Análisis estadístico

El conjunto de datos se dividió en las variables predictoras X (las primeras 187 columnas) y la variable objetivo y (la última columna). Después, X e y se dividieron en conjunto de entrenamiento (80 % de los registros) y conjunto de prueba (20 % de los registros) de manera estratificada para conservar la proporción de clases en cada conjunto.

Al aplicar estadística descriptiva a las variables predictoras, se observó que en varias columnas, el valor mínimo es 0 y el valor máximo 1, dando indicio de un escalamiento min-max en el intervalo [0,1]. De igual forma, al final hay columnas en donde la mayoría de los datos son 0, indicando asimetría positiva por los pocos outliers que no sean dicho valor. Entre más al centro estén las columnas, tienen valores más distribuidos con media alejada de 0 y desviación estándar mayor, pues se va llegando a la zona pico del ECG, en donde se da la mayor actividad ventricular. En la variable objetivo, al tener un mínimo de 0 y un máximo de 1 sin valores medios, se trata de un problema de clasificación binaria, pero al tener una media de 0.72, indica que aproximadamente el 72 % de los latidos son anormales, indicando un ligero desbalance en los datos.

Para ello se balancean las clases en los datos de entrenamiento para mejorar la diversidad geométrica y topológica de cada clase, permitiendo que el TDA capture estructuras más representativas de ambas clases, haciendo que el clasificador no tenga sesgo, logrando que se tenga una distribución más cercana al 50/50, aunque esto implique crear más registros de la clase que era minoría. Esto se hizo con la técnica de SMOTE-ENN, siendo una técnica híbrida de SMOTE (Synthetic Minority Over-sampling Technique), que genera muestras sintéticas de la clase minoritaria interpolando entre instancias vecinas dentro de esa clase (sobremuestreo), seguido de ENN (Edited Nearest Neighbours), donde se examina cada instancia y se eliminan aquellas que no coinciden con la mayoría de sus k vecinos más cercanos, eliminando así ruido y datos atípicos (submuestreo).

III-C. Algoritmos de TDA

Se utilizaron los conjuntos de entrenamiento para ajustar los modelos. Primero, con las variables predictoras de entrenamiento (X_train), se utilizó el modelo de Sliding Window Embedding para mapear cada latido a R^n, buscando $n \geq 1$, pues cada latido es una serie de tiempo en R, en donde se buscó el retardo τ óptimo utilizando la Información Mutua Promedio (AMI), seleccionando el primer mínimo local de la función AMI, obteniendo así $\tau = 1$. Se buscó también la dimensión de incrustación n óptima, mediante Descomposición en Valores Singulares (SVD), construyendo embeddings crecientes y calculando la proporción de varianza explicada por el menor valor singular, cuando esta cae por debajo de un umbral, se determina la n adecuada, obteniendo así $n = 4$.

Posteriormente, se calculan los diagramas de persistencia de cada latido embebido ahora en R^n, pero se consideran solamente las componentes conexas y los 1-huecos, debido a que calcular huecos de dimensión superior se vuelve computacionalmente costoso. Luego, se calculan los paisajes de persistencia a través de los diagramas de persistencia, obteniendo así el persistent Landscape de cada latido. Se consideraron 7 capas y una resolución de 300 valores.

Se realizó lo mismo con las variables predictoras de prueba (X_test), se realizaó la incrustación de Sliding Window con los mismos parámetros óptimos obtenidos, se calcularon los diagramas de persistencia y los persistent Landscape con los mismos parámetros considerados.

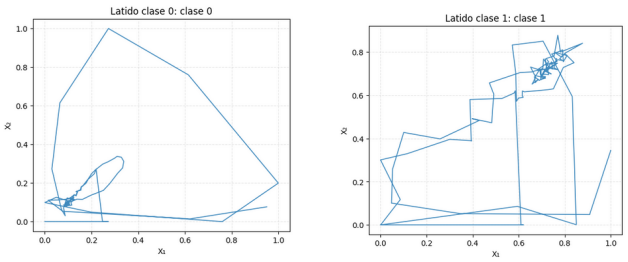
III-D. Algoritmos de ML

Posterior a los algoritmos de TDA, se realizaron algoritmos de Machine Learning (ML), donde se eligió el bosque aleatorio (RF), pues al promediar las predicciones de múltiples árboles de decisión se tiene una excelente capacidad de generalización y son muy eficientes al manejar datos de alta dimensionalidad, además de su fácil interpretabilidad y poco ajuste de hiperparámetros. El bosque aleatorio se entrenó utilizando los paisajes de entrenamiento de X_train, junto con y_train, para obtener los resultados finales de clasificación después de utilizar los algoritmos topológicos. De igual forma, se entrenó otro RF con los mismos hiperparámetros, pero utilizando los datos de X_train sin aplicarles los algoritmos de TDA, para comparar las diferencias.

IV. Resultados

Al utilizar Sliding Window Embedding con $n = 2$ y $\tau = 2$ se podían hacer gráficas en las que se podían observar diferencias cualitativas claras entre las trayectorias de los latidos normales (clase 0) y los anormales (clase 1). En los latidos de clase 0 (figura 1a), las curvas son más regulares y cerradas, con trayectorias más ”predecibles”que los latidos de clase 1 (figura 1b), donde se observa mayor dispersión y trayectorias más caóticas, es probable que los latidos de clase 1 tengan más ciclos significativos, de manera que el modelo pudo haberse aprovechado de ello.

Se realizó una Grid Search para encontrar los mejores hiperparámetros para ajustar el RF, de manera que se den los mejores resultados. Esto se realizó con los paisajes de



(a) Latido normal (clase 0). **(b)** Latido anormal (clase 1).

Figura 1: Trayectorias en el espacio de fases para latidos normales (a) y anormales (b).

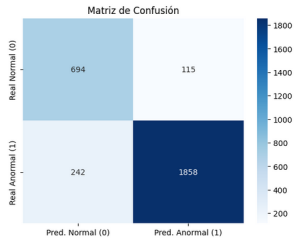
persistencia de X_train y con y_train. Obteniendo así los siguientes hiperparámetros:

n_estimators	Número de árboles en el bosque (150).
max_depth	Profundidad máxima de cada árbol (15).
min_samples_split	Mínimo de muestras necesarias para que un nodo se divida (5).
max_features	Proporción o número de variables consideradas en cada división (0.33).
random_state	Semilla para el generador aleatorio, garantiza reproducibilidad. En todos los algoritmos fue 69.
n_jobs	Número de núcleos CPU a usar en paralelo (‘-1‘ = todos).

Se obtuvo un AUC de 0.941442698216493, lo que indica que el modelo tiene una excelente capacidad de discriminación entre las dos clases. En el reporte de clasificación 2a se puede observar que el modelo alcanza un accuracy de 0.88, significando que acierta el 88 % de los latidos. El modelo detecta el 88 % de las arritmias reales con una precisión del 94 % y recupera correctamente el 86 % de los latidos sanos. En la matriz de confusión 2b se observa que la mayoría de los errores fueron falsos negativos (242), pues había algunas gráficas que incluso a simple vista se parecían mucho entre sí. Esto sugiere que las características topológicas capturan diferencias morfológicas relevantes.

	precision	recall	f1-score	support
0	0.74	0.86	0.80	809
1	0.94	0.88	0.91	2100
accuracy			0.88	2909
macro avg	0.84	0.87	0.85	2909
weighted avg	0.89	0.88	0.88	2909

(a) Reporte de clasificación.



(b) Matriz de confusión.

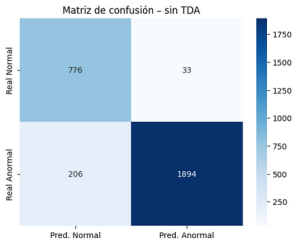
Figura 2: Métricas de la clasificación con TDA.

De manera similar, con el RF entrenado solamente con X_train y y_train sin TDA, se obtuvo un ROC de 0.9823394549414326 que indica una discriminación muy alta entre los latidos normales y anormales. El reporte de clasificación 3a muestra un accuracy de 0.92, además de que se identifican el 96 % de los latidos sanos con una precisión de 79 % y el 90 % de las arritmias con una precisión del 98 %. La

matriz de confusión 3b muestra que la mayoría de los errores son falsos negativos (206).

	precision	recall	f1-score	support
0	0.79	0.96	0.87	809
1	0.98	0.90	0.94	2100
accuracy			0.92	2909
macro avg	0.89	0.93	0.90	2909
weighted avg	0.93	0.92	0.92	2909

(a) Reporte de clasificación.



(b) Matriz de confusión.

Figura 3: Métricas de la clasificación sin TDA.

V. Conclusión

Aunque las métricas finales del Random Forest con descriptores topológicos quedaron ligeramente por debajo de las obtenidas sin TDA, los resultados confirman que la incorporación de Persistence Landscapes aporta una descripción robusta de la morfología de los latidos. Particularmente, las características topológicas pueden tener una alta capacidad para capturar patrones cíclicos y rupturas en la periodicidad de los datos, pudiendo complementar las señales crudas y mejorar así la interpretabilidad clínica.

El modelo TDA alcanzó un excelente equilibrio entre detectar arritmias y limitar falsos positivos, minimizando preocupaciones innecesarias, sin embargo, los resultados sugieren optimizaciones posibles, como mejorar los parámetros del embebido tomando un rango superior, tomar en cuenta los n-agujeros en el diagrama de persistencia (n¿1), probar esquemas de vectorización alternativos como imágenes de persistencia, buscar unos mejores hiperparámetros del bosque aleatorio, probar con otro algoritmo de aprendizaje automático, o incluso con redes neuronales, entre otras sugerencias.

Aunque el enfoque topológico no mostró una superación completa al baseline, logró ser una vía prometedora para enriquecer la representación de señales ECG con potencial para sistemas de apoyo al diagnóstico que requieran precisión y explicabilidad.

Referencias

Cascino, T., & Shea, M. J. (2023, diciembre). *Electrocardiografía*. Manual MSD. Consultado el 1 de mayo de 2025, desde https://www.merckmanuals.com/es-us/professional/trastornos-cardiovasculares/pruebas-y-procedimientos-cardiovasculares/electrocardiograf%C3%ADa

Chung, Y.-M., Hu, C.-L., Lo, Y.-L., & Wu, H.-T. (2021). A persistent homology approach to heart rate variability analysis with an application to sleep-wake classification. *Frontiers in Physiology*, *12*, 637684. https://doi.org/10.3389/fphys.2021.637684

de Chazal, P., O'Dwyer, M., & Reilly, R. B. (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, *51*(7), 1196-1206. https://doi.org/10.1109/TBME.2004.827359

Fazeli, S. (2018, mayo). *ECG Heartbeat Categorization Dataset*. Kaggle. Consultado el 1 de mayo de 2025, desde https://www.kaggle.com/datasets/shayanfazeli/heartbeat

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

Hernández-Lemus, E., Miramontes, P., & Martínez-García, M. (2024). Topological data analysis in cardiovascular signals: An overview. *Entropy*, *26*(1), 67. https://doi.org/10.3390/e26010067

Perea, J. A., & Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, *15*(3), 799-838. https://doi.org/10.1007/s10208-014-9206-z

PhysioNet. (2004, septiembre). *PTB Diagnostic ECG Database (versión 1.0.0)*. PhysioNet. Consultado el 1 de mayo de 2025, desde https://www.physionet.org/content/ptbdb/1.0.0/

Anexos

- Código en Google Colab: https://colab.research.google.com/drive/1--n71X6QPnn-4H1ew_5uI27EOO44JsaV?usp=sharing